A Multimodal Scheduler for Synchronized Humanoid Robot Gesture and Speech

Maha Salem[1], Stefan Kopp[2], Ipke Wachsmuth[3], and Frank Joublin[4]

[1] Research Institute for Cognition and Robotics, Bielefeld University, Germany,
msalem@cor-lab.uni-bielefeld.de
[2] Sociable Agents Group, Bielefeld University, Germany,
skopp@techfak.uni-bielefeld.de
[3] Artificial Intelligence Group, Bielefeld University, Germany,
ipke@techfak.uni-bielefeld.de
[4] Honda Research Institute Europe, Offenbach, Germany,
frank.joublin@honda-ri.de

1 Abstract

In order to engage in natural and fluent human-robot interaction, humanoid robot companions must be able to produce speech-accompanying non-verbal behavior including hand and arm gestures. In human communication, gestures are considered an integral part of the human thinking process. Accordingly, they are found to be finely synchronized with the accompanying linguistic affiliate [1]. Many researchers have emphasized the importance of this temporal synchrony in terms of co-expressiveness (e.g. [2], [3]). However, for a humanoid robot required to generate speech and gesture, an appropriate synchronization of the two modalities still poses a major challenge. In many existing approaches used for virtual conversational agents or robotic platforms, synchronization of different modalities is either achieved only approximately or by solely adapting one modality to the other, e.g. by adjusting gesture speed to the timing of running speech. Given the limitations of robotic platforms, e.g. motor velocity limits, these approaches turn out to be insufficient when a fine synchronization of speech and gesture is a fundamental necessity for fluent human-robot interaction.

We present a multimodal scheduler that is capable of synchronizing expressive hand and arm gestures with speech for the Honda humanoid robot. Since the challenge of multimodal behavior realization has already been tackled in various ways within the domain of virtual agents, our approach exploits the experiences gained from the development of a speech-gesture production model for embodied conversational agents. In particular, we build on the Articulated Communicator Engine (ACE), which is one of the most sophisticated multimodal schedulers and behavior realizers in that it replaces the use of lexicons of canned behaviors with a real-time production of flexibly planned behavior representations [4]. The implementation of the interface that couples ACE with the perceptuo-motor system of the Honda robot and which is now used as an underlying action generation framework for the humanoid is described in [5]. An outline of the implemented robot control architecture is shown in Figure 1.

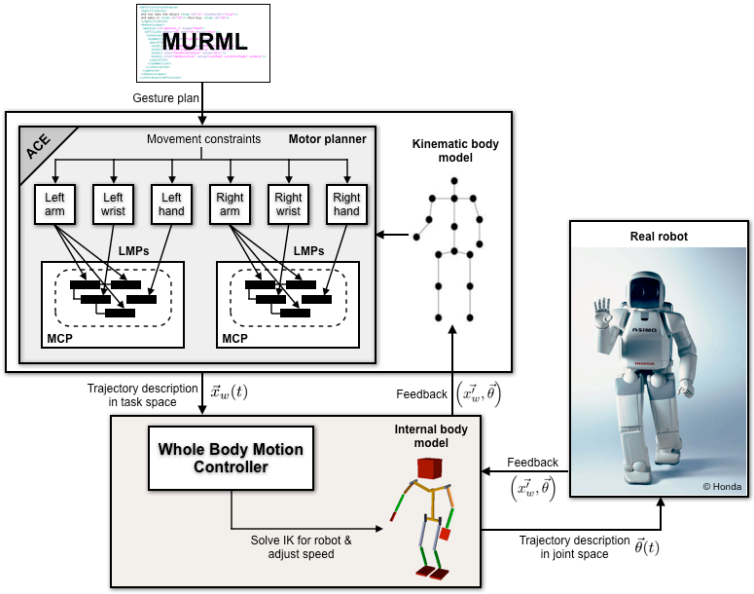

Fig. 1. Outline of the underlying robot control architecture.

The scheduler presented in this abstract is an extended and improved version of the ACE scheduler originally developed for a virtual agent application, since it was lacking the two following major functionalities when used on a robotic platform:

1. **Accurate prediction function for movement timing**: Being designed for an animated virtual character, the forward model implemented in ACE to estimate the time needed for the agent to perform a body movement is based on a simple estimation function (Fitt's law). This was sufficient in the virtual agent environment, however, has proven inadequate when the scheduler was used for action generation with a physical humanoid robot.
2. **Cross-modal adaptation mechanisms within a multimodal chunk**: In ACE, within a chunk of speech-gesture production (see definition in [4]), gesture speed is adjusted to the timing of running speech. This is possible in a virtual agent application, since there are no real physical limitations in the animation of body movements. Given a robotic platform, however, motor velocity limits and other physical restrictions constrain the execution of arm and hand gestures with arbitrary speed.

Essentially, our proposed multimodal scheduler incorporates a forward model to predict an estimate of the preparation time required by the robotic body for a gesture prior to the actual gesture stroke taking place. For this, an internal simulation of the designated arm movement is performed during the movement

planning phase using the robot's whole body motion controller software [6]. Note that the same controller is also used to subsequently generate the actual movement of the robot. Despite the fairly accurate prediction-based timing estimation, the actual execution and timing of multimodal utterances might deviate from the prediction. For this reason, an on-line adjustment of the synchronization process is required once a certain threshold value of deviation is exceeded. This is achieved by mutually adapting the two modalities reactively to one another. In this way, our extended scheduler allows for cross-modal adaptation not only between, but also within chunks. Scheduling, generation and continuous synchronization of gesture and speech are flexibly conducted at run-time. In the following, the implementation of the features extending the original ACE scheduler is described in more detail.

1.1 Implementation of extended features

The ACE scheduler augments the classical two-phase 'planning-execution' procedure with additional phases of the speech-gesture production process. However, for the sake of simplicity, we use the classical fragmentation to illustrate the extended features. In the proposed scheduler, the two phases operate as follows (new additions or changes in the model compared to the original ACE scheduler are marked with *):

Phase 1: Planning

1. a) Phonological encoding
   b) Determine timing information for speech
   c) Generate speech output in two files (1. speech before affiliate onset // 2. affiliate and rest of speech)*
2. a) Movement planning
   b) Predict gesture preparation time based on forward model using robot's whole body motion controller*
   c) Set start times for speech and gesture based on comparison of timing information for both modalities*: If duration of speech before affiliate onset is longer than gesture preparation time before stroke onset, then start with speech output; otherwise start with gesture execution.

Phase 2: Execution

1. Start with speech-gesture production as scheduled based on timing information derived in 2c) of planning phase*
2. Constantly check variance between target and actual arm position utilizing afferent feedback from robot*
3. If variance exceeds defined threshold, adjust the synchronization process through reactive mutual adaptation:*
   - If gesture is too slow, then pause speech before affiliate onset (i.e. wait with playback of second sound-file) until variance is below threshold
   - If gesture is too fast, then execute pre-stroke hold (unlikely to happen)

Figure 2 illustrates the two phases of the extended ACE scheduler.

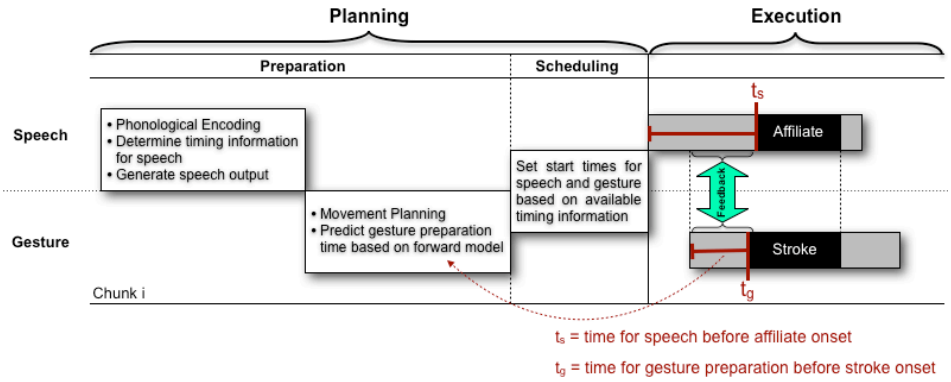

Fig. 2. Illustration of the 'planning-execution' procedure of the extended scheduler.

Conclusion. The implementation of the proposed scheduler enables the Honda humanoid robot to plan, generate and continuously synchronize gesture and speech at run-time. The extended scheduler incorporates a predictive model for body movements and cross-modal adaptation mechanisms. It thus represents a more flexible and natural way to realize multimodal behavior for robots and other artificial communicators.

Acknowledgments. The work described is supported by the Honda Research Institute Europe.

References

1. McNeill, D., Quek, F., McCullough, K.-E., Duncan, S., Furuyama, N., Bryll, R., Ma, X.-F., Ansari., R.: Dynamic Imagery in Speech and Gesture. In: B. Granström, D. House and I. Karlsson (Eds.) Multimodality in Language and Speech Systems. Dordrecht, The Netherlands: Kluwer Academic Pulblishers (2002)
2. McNeill, D., Bertenthal, B., Cole, J., Gallagher, S.: Gesture-first, but no Gestures? Behavioral and Brain Sciences 28 (2):138–139 (2005)
3. Habets, B., Kita, S., Shao, Z., Özyurek, A., Hagoort, P.: The Role of Synchrony and Ambiguity in Speech: Gesture Integration During Comprehension. Journal of Cognitive Neuroscience. Advance Online Publication (2010)
4. Kopp, S., Wachsmuth, I.: Synthesizing Multimodal Utterances for Conversational Agents. Computer Animation and Virtual Worlds 15(1):39–52 (2004)
5. Salem, M., Kopp, S., Wachsmuth, I., Joublin, F.: Towards an Integrated Model of Speech and Gesture Production for Multi-Modal Robot Behavior. In: Proceedings of the 2010 IEEE International Symposium on Robot and Human Interactive Communication (2010)
6. Gienger, M., Janen, H., Goerick, C.: Task-Oriented Whole Body Motion for Humanoid Robots. In: Proceedings of the IEEE-RAS International Conference on Humanoid Robots, Tsukuba, Japan (2005)